# ISOPB 2017

## Agrigenomics in the breeder's toolbox: latest advances towards an optimal implementation of genomic selection in oil palm

Jacob, F[1]; Cros, D[2]; Cochard, B[1] and Durand-Gasselin, T[1]

[1]PalmElit S.A.S, 2214 Boulevard de la Lironde, 34980 Montferrier-sur-Lez, France
[2]CIRAD, UMR AGAP (Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit), 34398 Montpellier, France

## Abstract

PalmElit implements the genetic improvement and marketing programs for CIRAD® oil palm seeds. The commercial seeds embody 80 years of genetic improvement work undertaken by IRHO, CIRAD and PalmElit in conjunction with several partners of excellence located on each of the continents where oil palm is grown. An increase of more than 60% in oil yields was achieved since 1960. This result illustrates the efficiency of the recurrent reciprocal selection (RRS) underlying the conducted breeding program. So far, assessment of parental breeding values has largely relied on progeny testing, which is an efficient but time- and money-consuming step within the RRS scheme. With the recent development of oil palm genomic resources, genomic selection (GS) appears as an attractive strategy to increase the efficiency of oil palm breeding programs. On a theoretical point of view, GS has the potential to increase the rate of genetic gain by shortening the breeding cycle and/or increasing the selection intensity.

PalmElit, together with its research partner CIRAD, has been leading research for nearly 10 years in order to develop and assess the implementation of GS in oil palm breeding. Some of the key achievements have been shared with the scientific community since 2015 (Cros et al., 2015a, 2015b, 2017a; Marchal et al., 2016) which corroborate the potential of GS in terms of increased genetic gain. Further research is still ongoing to answer the simple -but critical- question: what is the optimal use of GS in terms of genetic gain vs time- and cost-efficiency? In this paper, following a brief review on the GS history and key concepts, we present our latest results which address critical aspects such as prediction accuracy and optimal use of GS within breeding schemes. We extend and discuss our conclusions in light of the literature available in oil palm and other crop species. Finally, we summarize the perspectives and challenges for successful implementation of GS in oil palm.

## Introduction

### *Pros and cons of the classical recurrent reciprocal selection*

Oil palm varieties typically consists in *tenera* hybrid crosses between heterotic group A (mostly Deli origin, *dura* palms) and group B (mostly African origins, *pisifera* palms). Selection and breeding among the parental populations usually relies on progeny testing since hybrid performances might not be accurately predicted based on parental performances (Corley and Tinker, 2015a). In order to achieve an efficient and sustained improvement of its commercial hybrids, PalmElit employs a

recurrent reciprocal selection (RRS) strategy for both group A and B parental populations (Baudouin et al., 1997). This strategy aims at improving the general combining abilities (GCA) of the parental population along the successive breeding cycles. Pros and cons of the RRS in oil palm have been already debated (Corley and Tinker, 2015a). According to Gallais (Gallais and Poly, 1990), the main advantages of recurrent selection are:

- increasing the frequency of genes and associations favoring the type of variety to be developed
- enabling effective recombination, hence highly effective multi-trait breeding
- preventing an over-rapid loss of variation, provided it is carried out correctly
- partially fixing heterosis
- ensuring continuous, long-term progress
- providing outputs directly applicable for varietal creation

When RRS is applied in oil palm, one breeding cycle extends over a long period of time (~20 years) in contrast with some annual crops (e.g. 3 months in rice). Despite this long cycling time, a high genetic gain rate has been achieved since 1960 (~+1%/year for yield, Durand-Gasselin et al., 2010), highlighting the potential of oil palm in terms of genetic improvement. The main limitation in terms of cost- and time-efficiency relates to estimation of the parental GCA and hybrid values since it traditionally requires progeny testing for each parent. Thus, techniques allowing faster and/or cheaper GCA or hybrid value estimation could greatly improve oil palm breeding, including RRS.

***New tools of the agrigenomics era: marker assisted selection and genomic selection***

As more genetic and genomic resources become available for oil palm, new breeding tools become available such as marker assisted selection (MAS, reviewed for crops in Collard and Mackill, 2008). In MAS, molecular marker data can be used to predict phenotype(s), based on known association between the chosen marker(s) and phenotype(s). Marker-phenotype associations can be identified using approaches such as quantitative trait loci (QTL) mapping. In that case, markers linked with the strongest QTLs can be selected and used for predicting the associated phenotype. This selection method can be efficient provided that:

- MAS is faster and/or cheaper than the conventional phenotypic screening
- QTLs are accurately identified (appropriate experimental design to guarantee a high detection power and to limit the risk of false positives and of QTL effect overestimation)
- linkage between markers and QTLs is strong
- association between marker(s) and the phenotype(s) is conserved in the population and the environment where the selection will be carried out
- a limited number of QTLs accounts for a sufficient part of the phenotypic variation (e.g. the trait is essentially mono- or oligogenic)

The latter point defines one major drawback of the classical MAS strategy since many agronomic traits are quantitative and thus likely influenced by a large number of loci. Genomic selection (GS) was developed as a specific case of MAS designed for quantitative traits (Meuwissen et al., 2001). In genomic selection, individuals are genotyped over a dense set of genome-wide markers that can ideally account for all QTLs in the genome. Based on marker data, a genomic estimated breeding value (GEBV, with BV and GCA generally linked by BV = 2*GCA) can be assigned to each genotyped individual provided that the model was calibrated using an appropriate training set (TS) which combines genotypic and phenotypic data for the trait(s) of interest.

GS was first developed and implemented for cattle breeding and has later found its way to plant breeding. Publication trends clearly illustrate the research expansion for GS in plants as of 2009

(Figure 1). Despite the amount of research conducted, and the growing evidence for its potential in hybrid breeding (Marulanda et al., 2016; Zhao et al., 2015), practical implementation of GS has remained limited to a few species including wheat, maize, rye, pines, cassava, and recently oil palm (Cros et al., 2017a; Kwong et al., 2017).

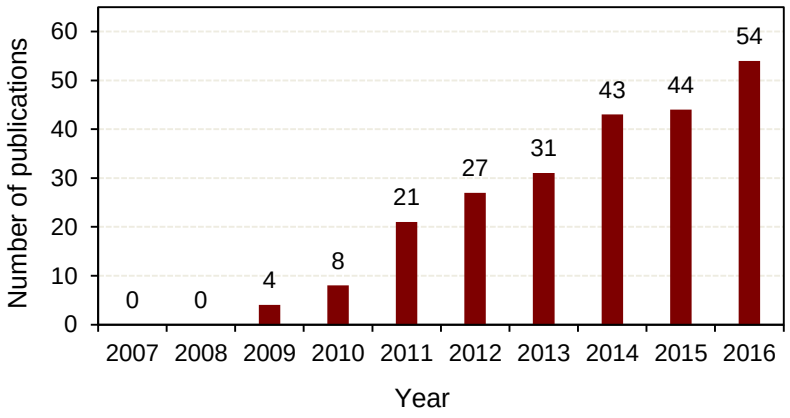



Figure 1: Publication trends for GS in plants. Publications trends were estimated by counting the number of publications in the plant field that contains "genomic selection" in the title and are referenced on Google Scholar.

***Potential of GS in oil palm breeding***

As evoked earlier, GS could improve many aspects of the oil palm breeding programs:

- the estimation of the value of hybrid crosses which have not been phenotyped. In that respect, GS can directly support the identification and selection of commercial hybrid with higher agronomic value
- the estimation of the GCA of individuals among the germplasm. In that case, GS can assist the process of recombination within the germplasm to increase the genetic value of the parental population
- the duration of the breeding cycle (reduction) by replacing part of, or the entire phenotyping process
- the selection intensity (increase) for both hybrid crosses and parental populations by including individuals for which only genotypic data is available

In the following article, we review the latest results which address critical aspects such as prediction accuracy and optimal use of GS within breeding schemes. We extend and discuss our conclusions in light of the literature available in oil palm and other crop species. Finally, we summarize the perspectives and challenges for successful implementation of GS in oil palm.

## Current status for GS in oil palm

### *From classical breeding to genomic selection at PalmElit: past, present, and future*

In the past, for CIRAD® germplasm, GCA estimation was based on appropriate statistical analysis of genetic trials with complex experimental designs (e.g. incomplete and unbalanced factorial designs). Lately, we implemented a pedigree-based best linear unbiased prediction (T-BLUP) approach to improve parental GCA estimation in the context of these complex genetic trial designs. By borrowing information from the pedigree (under the form of a kinship matrix), pedigree-based BLUP could also estimate GCA of individuals which are not tested but are related to progeny-tested individuals (P-BLUP, Figure 2A-B). This example illustrates how appropriate statistics can estimate GCA of untested individuals, provided that a suitable training set (TS) is available for calibrating the model.

Study on seven yield components indicated that GCA prediction accuracy using P-BLUP is intermediate to high depending on the trait and the heterotic group considered (ranging from 0.22 to 0.82, Table 1, Cros et al., 2017a). However, this approach requires accurate knowledge of the germplasm pedigree (Corley and Tinker, 2015a) and cannot account for Mendelian sampling. This is illustrated by the fact that pedigree information cannot discriminate individuals within full-sib families although these have distinct genotypes as a result of Mendelian segregation (Figure 2A-B). Thus, P-BLUP-based GCA estimation is not suitable for intra-family selection. To overcome this limitation, we decided to test whether GS could perform better than P-BLUP (Figure 2). For genome-based predictions, we used a similar BLUP model, that we designated as G-BLUP. The G-BLUP method was successfully applied for hybrid prediction in various species, including maize, soybean, rice, triticale and sunflower (Zhao et al., 2015). Moreover, a previous study of Cros et al. indicated that this model performs similarly to several other tested models when applied on empirical oil palm data (Cros et al., 2015a).

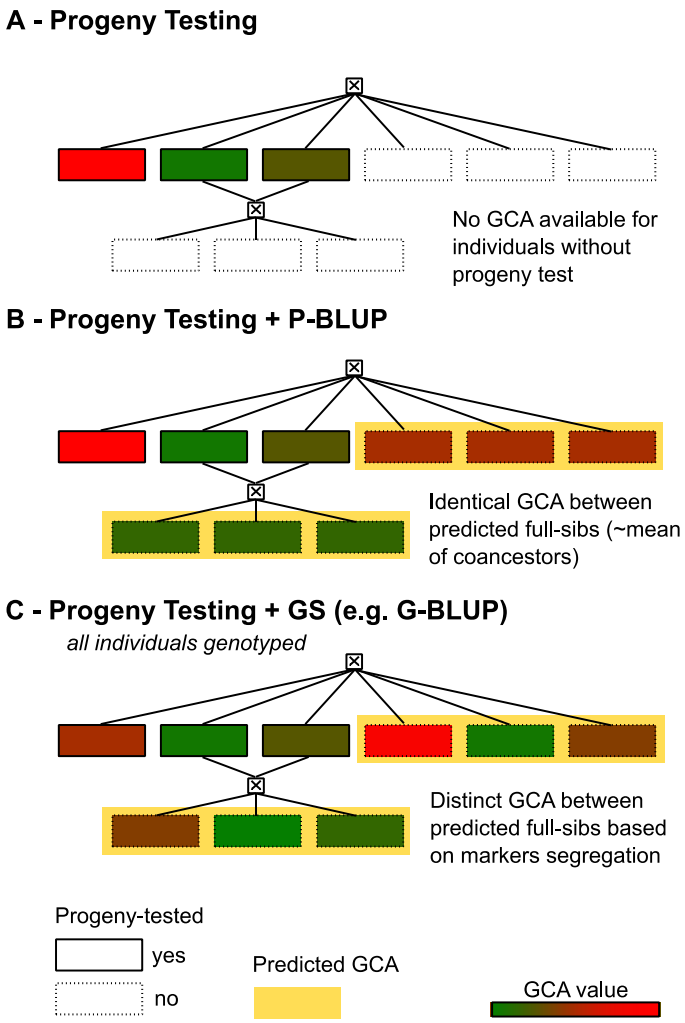



Figure 2: Comparison of GCA calculation methods based on a simple case.

The general model used for GCA prediction can be written as follows:

$$Y = X\beta + Zb + Z_A g_A + Z_B g_B + Z_D s_{AB} + e$$

where $Y$ is the vector of the phenotypes of the hybrid individuals, $\beta$ and $b$ are the vectors of fixed and random effects due to the experimental design, respectively, $X$ and $Z$ their associated incidence matrices, $g_A$ and $g_B$ are the vectors of GCA (additive effects) of A and B parents, respectively, $s_{AB}$ is the vector of SCA (dominance effects) of crosses, $Z_A$, $Z_B$ and $Z_D$ their incidence matrices and $e$ is the vector of residual effects. Covariance definition for GCAs defines the main difference between P-BLUP and G-BLUP. For P-BLUP, the covariance is derived from genealogical relationships (pedigree information) whereas for G-BLUP, it is derived from genomic relationships (marker data).

A training set (TS) corresponding to ~500 crosses from 150 A parents and 156 B parents grown in one site in Indonesia was used to predict values for a validation set (VS) of ~200 crosses from 67 A parents and 42 B parents grown in another location in Indonesia (for details, see Cros et al., 2017a). The parents of both TS and VS were genotyped using genotyping-by-sequencing (GBS) which produced >5000 high quality SNPs suitable for GS. The hybrid crosses were phenotyped but not

genotyped. Comparison of the prediction accuracies between P-BLUP and G-BLUP indicated that G-BLUP can perform better than P-BLUP depending on the group and the trait (Table 1 and Figure 3). This observation holds true for parental GCA and hybrid value prediction. The best improvement was obtained for FFB.

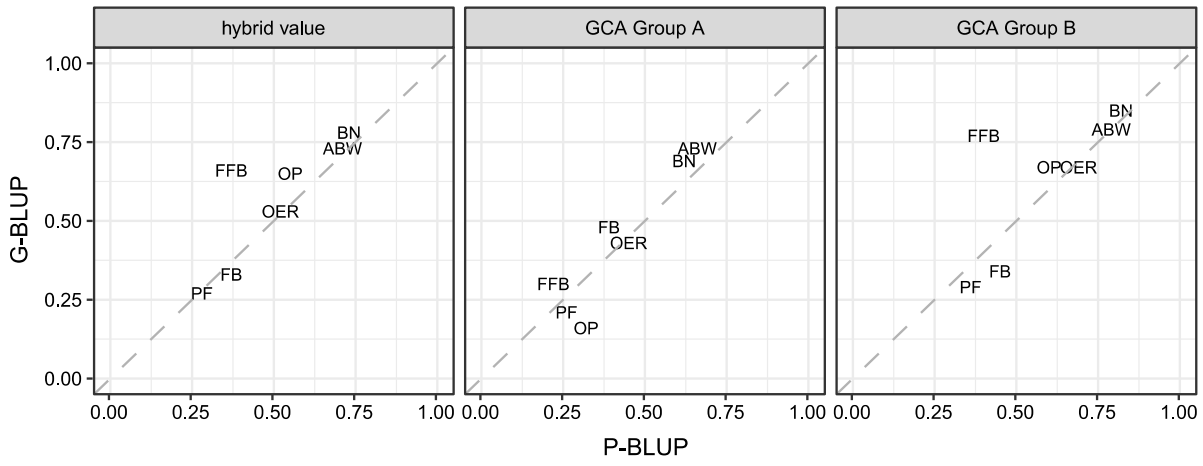



Figure 3: Correlation between P-BLUP and G-BLUP prediction accuracies. The plots are based on the same data as shown in Table 1.

Table 1: Accuracy of P-BLUP and G-BLUP for yield components in the study of Cros et al. 2017a

| | Prediction accuracy across populations (sites) | | | | | |
|---|---|---|---|---|---|---|
| | Hybrid value | | GCA Group A | | GCA Group B | |
| Yield component | P-BLUP | G-BLUP | P-BLUP | G-BLUP | P-BLUP | G-BLUP |
| FFB | 0.37 | **0.66** | 0.22 | **0.30** | 0.40 | **0.77** |
| BN | 0.73 | **0.78** | 0.62 | **0.69** | 0.82 | **0.85** |
| ABW | 0.71 | **0.73** | 0.66 | **0.73** | 0.79 | 0.79 |
| FB | 0.37 | 0.33 | 0.39 | **0.48** | 0.45 | 0.34 |
| PF | 0.28 | 0.27 | 0.26 | 0.21 | 0.36 | 0.29 |
| OP | 0.55 | **0.65** | 0.32 | 0.16 | 0.60 | **0.67** |
| OER | 0.52 | **0.53** | 0.45 | 0.43 | 0.69 | **0.67** |

FFB: annual cumulative fresh fruit bunch, in kg
BN: annual cumulative bunch number
ABW: annual average bunch weight, in kg
FB: fruit-to-bunch ratio, in kg
PF: pulp-to-fruit ratio, in %
OP: oil-to-pulp ratio, in %
OER: oil extraction rate, in %
For G-BLUP, the accuracy corresponds to the accuracy obtained with the maximum number of SNPs
Bold: G-BLUP prediction accuracy higher than P-BLUP

A simulation study was performed to assess the potential gain when employing GS as a preselection step on FFB within the classical RRS scheme (Figure 4). For example, FFB in the top 100 hybrid crosses could be increased by ~11% when applying the preselection on a breeding population of 5000 A and 5000 B palms This example demonstrates that a simple preselection step using GS can already greatly improve the genetic gain in commercial hybrids.

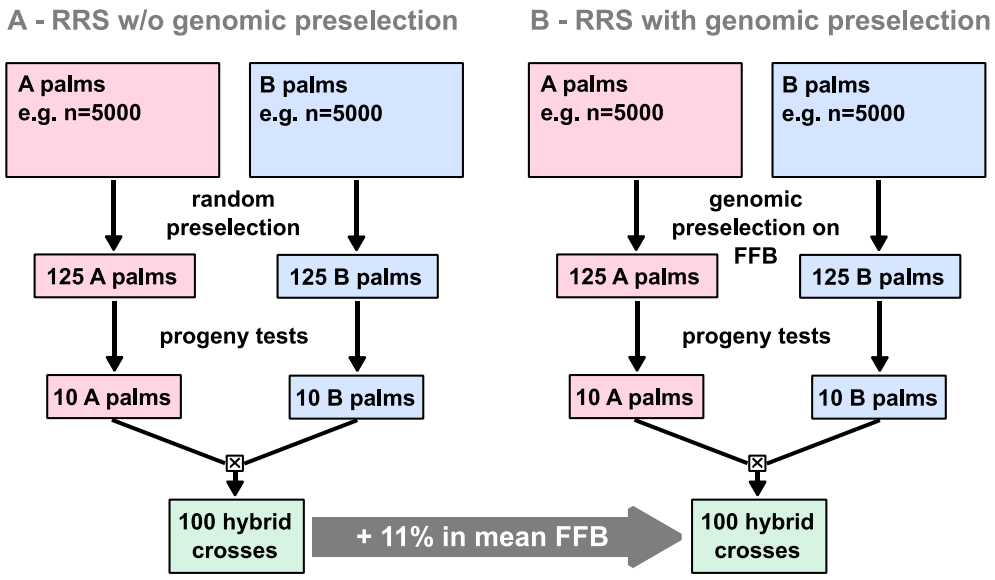



Figure 4: Representation of the simulation design used to estimate the genetic gain of genomic preselection on FBB (B) compared to a classical RRS scheme (A). The analysis is described in Cros et al. 2017a

***Overview of the research published by other entities***

Until now, very few studies have been published on GS applied to oil palm. We briefly summarize here what is published besides research conducted within PalmElit's network (Cros et al., 2015a, 2015b, 2017a; Marchal et al., 2016).

The first publication on GS in oil palm was presented by Wong and Bernardo (Wong and Bernardo, 2008). This work was conducted in association with Applied Agricultural Resources Sdn. Bhd (AAR). Based on simulated data for a small oil palm parental population derived from a single cross, Wong and Bernardo demonstrated the potential of genomic selection compared to phenotypic selection and QTL-based marker-assisted selection. The study also provided the first estimates of gain depending on parameters such as the size of the breeding population, the number of replications in phenotypic assays, and the heritability of the trait. The cost per unit gain and the time per unit gain were calculated to assess the efficiency of each breeding strategies conducted over 37-38 years (corresponding to 2 cycles of classical phenotypic selection or 4 cycles of marker-assisted or genomic selection). The improvement obtained with GS (up to +25% in the response to selection with a population size N=70, and cost per unit gain reduced by at least 26% compared to phenotypic selection) was mainly attributed to the shorter generation time when selection was based solely on genotypic data (6 years vs 19 years for a traditional selection cycle). This analysis also suggests that increasing the number of parental palms tested could be more efficient than increasing the number of replication in field tests. However, since this study was conducted with simulated data under specific assumptions, the results need to be validated with empirical data.

A recent study by Sime Darby reported interesting results related to the implementation of GS for early selection among commercial hybrid populations (Kwong et al., 2017). 1,218 commercial hybrids were genotyped and phenotyped for 6 production traits with varying heritability. The GS strategy

applied was to use part of the hybrid population as TS to predict the value of the other part (=VS). This study mainly focused on optimizing the marker set and statistical method to maximize the prediction accuracy while reducing the number of markers (potentially leading to reduced genotyping costs). The results of Kwong et al. are further discussed below.

## Implementation and optimization of GS in oil palm

### *Considerations for optimizing the use of GS*

Several studies provide evidence for the efficiency of genomic prediction for increasing the gain in agronomic traits (e.g. in oil palm: Cros et al., 2017a; Kwong et al., 2017). Still, there is a large gap between the experimental studies and GS routine implementation in breeding. Several facts can explain such discrepancy:

- Additional costs due to genotyping might render the method less profitable. Thus, GS implementation implies an accurate estimation of both gain per time and gain per costs.
- Optimal use of GS could require profound changes in the breeding scheme, with an impact on the traditional breeding practices.
- GS might not be the most optimal selection method for all agronomic traits.
- Commercialization of planting material selected solely on the basis of its genomic estimated value might be problematic due to the absence of phenotypic records to demonstrate its real agronomical value and check secondary traits which have not been selected for.

Many studies have focused on determining the optimal parameters for maximizing the gain while fewer also included cost considerations (e.g. Rajsic et al., 2016; Wong and Bernardo, 2008). Figure 5 summarizes some of the main parameters which affect GS gain and cost, some of which will be further discussed below.

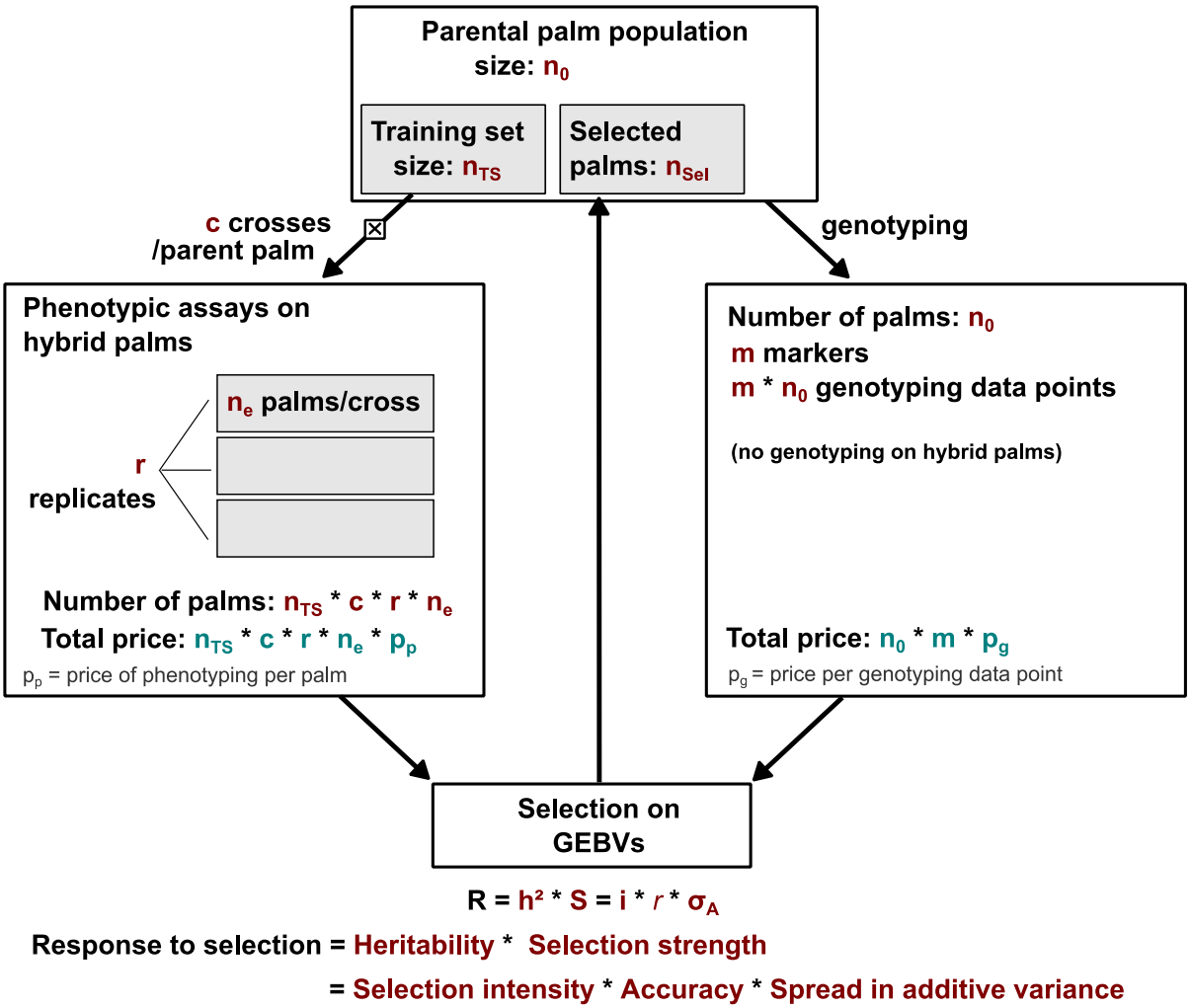



Figure 5: Parameters affecting the gain and cost of GS. Parameters affecting the gain and cost are indicated in brown and cyan respectively. The link between cost and the indicated parameters is direct and given by the formula indicating the total cost for both phenotyping (left) and genotyping (right). The link between the design parameters and the gain (response to selection) is indirect with $n_{Sel}$ and $n_0$ affecting i, and nearly all design parameters affecting the prediction accuracy *r*. Omitted here is the generation time, which impacts the gain rate and can be reduced by skipping part or all of the longest phenotyping assays.

❖ Key parameters for gain maximization

As shown in the formula used to estimate the response to selection (Figure 5), i (selection intensity) and *r* (selection accuracy) are key parameters which can be adjusted to optimize GS while $\sigma_A$ is mainly an intrinsic genetic feature of the trait under selection within the considered breeding population. Overall, increasing the size and diversity within the breeding population, and increasing the prediction accuracy positively contribute to the GS efficiency. The gain rate could also be increased by shortening the generation time, which would imply to decrease the time spent on phenotypic tests, e.g. by implementing progeny tests at lower frequency throughout the cycles.

❖ Cost minimization

It has been a general concern among breeders that GS would increase the breeding costs. Such concern should however be allayed by several facts:

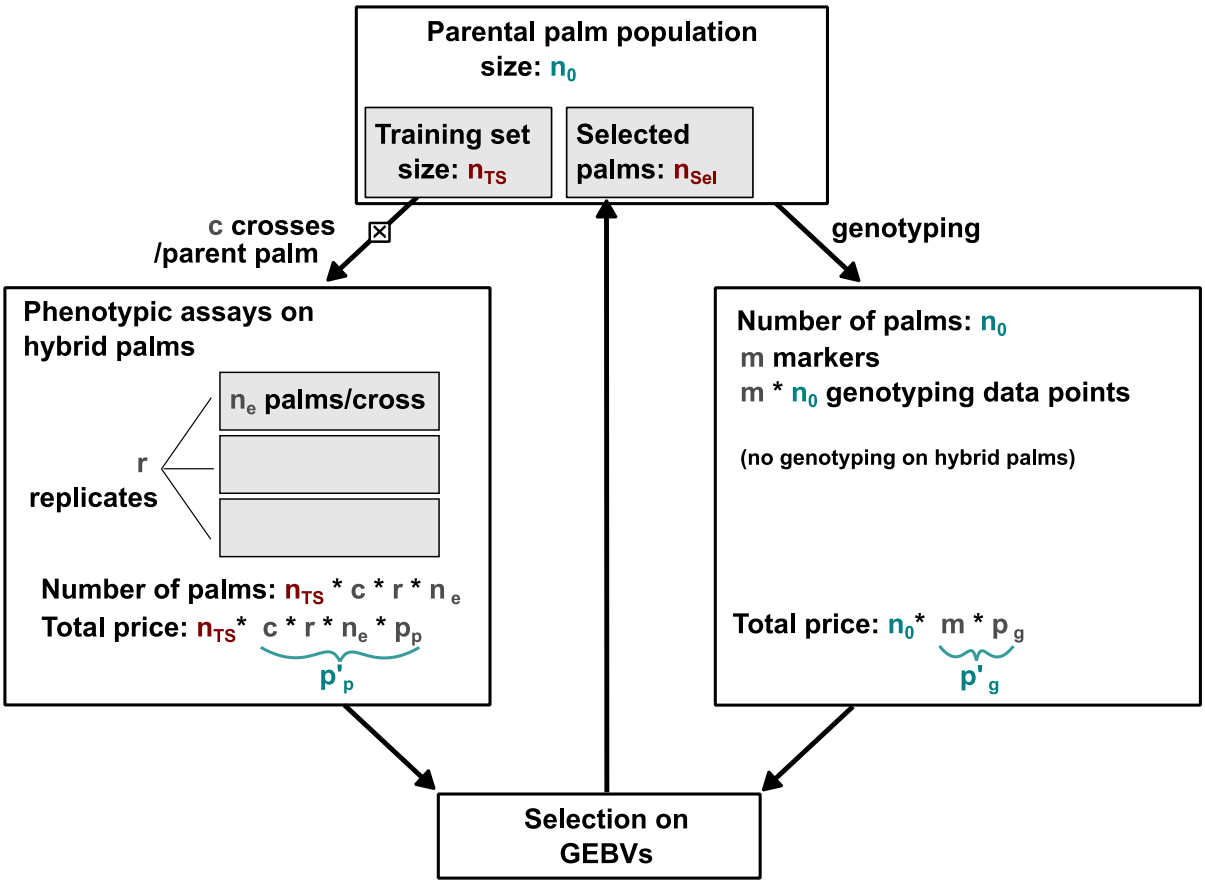
- Genotyping costs remains much smaller compared to phenotyping costs, and both technological progress and increasing labor costs will contribute to widen the gap between them in the future. To illustrate such difference, progeny testing in one of our on-going standard genetic trial in Nigeria costs 6,400-22,000€ per parental palm while

genotyping cost will likely be below 50€/palm for a large-scale implementation and a reasonable number of SNPs (examples of current estimates are ~50-60€/palm by GBS and ~200-300€/palm with the OP300K SNP). Thus, even with the highest genotyping cost (300€) and the lowest phenotyping costs (6,400€), the ratio of genotyping/phenotyping costs lays below 1/20.

- The cost increase related to GS can be compensated by a decrease in the cost of the phenotypic assays. For example, a similar selection accuracy could be achieved using a smaller but better designed training set (Rincent et al., 2017; Wolfe et al., 2017).
- Considering the current genotyping costs per individual applying GS for direct selection in commercial hybrids seems not economically viable, unless the selected hybrids can provide a sufficient return on investment (e.g. by cloning them). It is possible to focus the genotyping effort on the parental palms, allowing both parent and hybrid prediction, while reducing the genotyping costs (Cros et al., 2015a, 2015b, 2017a; Marchal et al., 2016) compared to strategies where hybrid individuals are also genotyped (Kwong et al., 2017).

❖ *In silico* breeding approach for GS design optimization

Based on the same scheme described in Figure 5, which illustrates a simple use of GS for breeding within a parental population, we tested the impact of several parameters on costs and genetic gain. The analysis design is described in Figure 6 and the results in Figure 7.




**Scenario 1: parental palm selection without GS**
**Scenario 2: parental palm selection with GS**

**Response to Scenario 2 ($R_2$) / Response to Scenario 1 ($R_1$)**

$$R_2/R_1 = i_2/i_1 * r_2/r_1$$

**Fixed parameters**
$n_{TS}$ = 100
$n_{Sel}$ = 25

**Varying parameters**
$n_0$ = 100 - 5000 = number of palms in the parental population
$p'_p = c * r * n_e * p_p$ = 6400 - 22000 € = phenotyping costs per parental palm
$p'_g = m * p_g$ = 50 - 300 € = genotyping costs per palm
$i_2$ is directly derived from $n_0$
$r_2/r_1$ = 0.8 - 1.8 (estimation derived from Cros et al. 2017)

Figure 6: Design used to test the influence of several parameters on GS gain and cost. Parameters with a unique fixed value are indicated in brown. Parameters with variable values are indicated in cyan, the range allowed is delimited by extreme values estimated based on empirical data. Parameters which are not included are in grey. Parental palm selection without GS (Scenario 1) consists in progeny-testing $n_{TS}$=100 parent palms and selecting the top $n_{Sel}$=25 palms. Parental palm selection with GS (Scenario 2) consists in progeny-testing $n_{TS}$=100 parent palms, estimating breeding values for $n_0$ breeding palms including the training set after having all of them genotyped, and selecting the top $n_{Sel}$=25 palms. C: number of crosses per tested palm. r: number of replicates for each tested cross. $n_e$: number of palms tested per cross. $p_p$: phenotyping cost per hybrid palm in the field. m: number of markers. $p_g$: cost per marker data point. $p'_p$ and $p'_g$: phenotyping and genotyping costs per breeding palm respectively. i: selection intensity. *r*: selection accuracy.

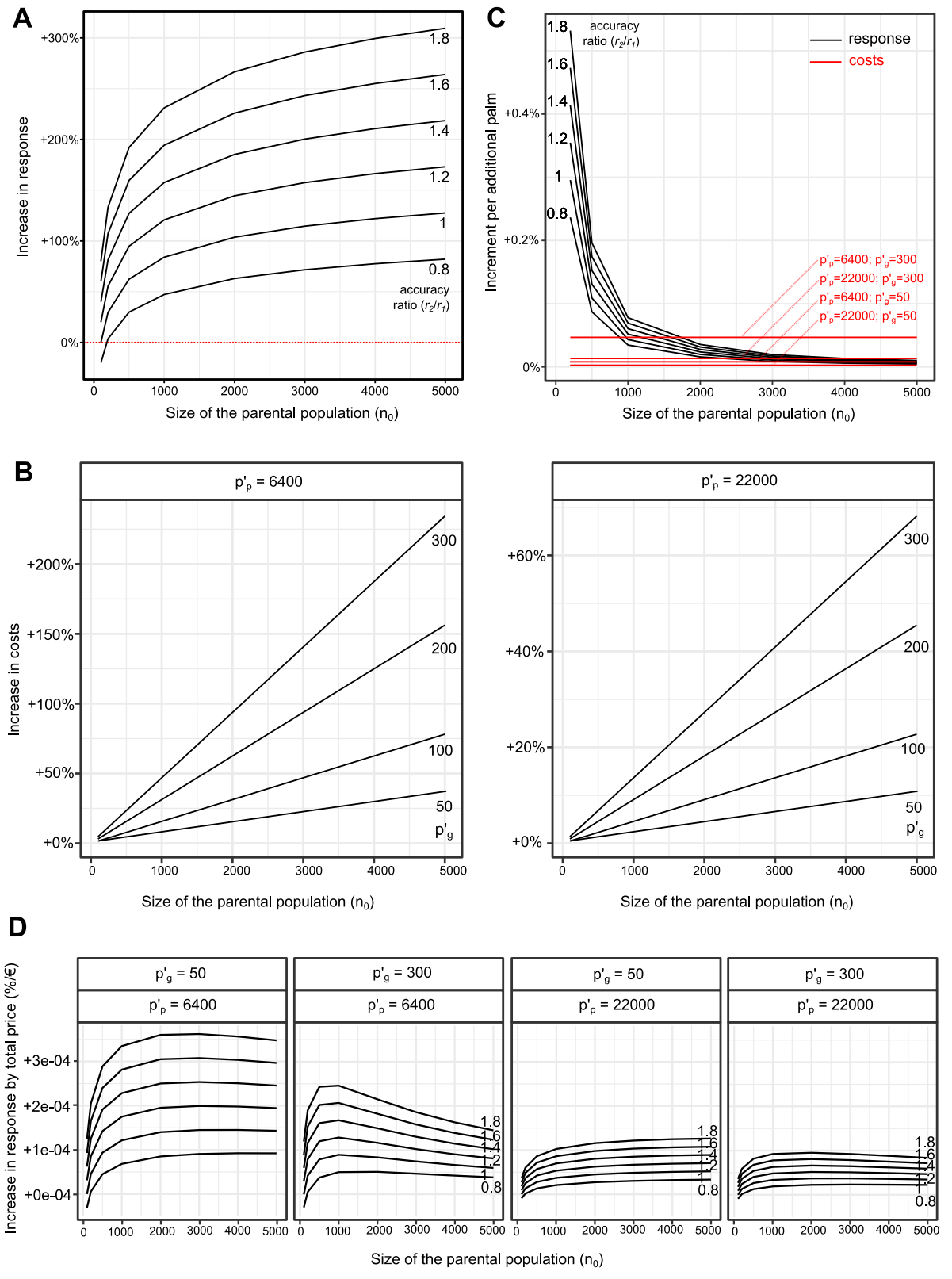



Figure 7: Influence of several parameters on GS gain and cost. The parameters tested are described in Figure 6: increase in response to GS as compared to phenotypic selection (($R_2$-$R_1$)/$R_1$), genomic vs phenotypic selection accuracy ratio ($r_2/r_1$), phenotyping and genotyping costs per breeding palm ($p'_p$ and $p'_g$ respectively), and size of the breeding population ($n_0$) **A.** Additional genetic gain when using GS as compared to selection based on traditional progeny testing for different accuracy ratios between GS and the traditional method, and depending on the population size. **B.** Increase in costs depending on the population size, phenotyping costs, and genotyping costs. **C.** Cost and gain increment when increasing the population size, depending on the accuracy ratio, phenotyping costs, and genotyping costs. **D.** Increase in genetic gain per total cost, depending on the population size, phenotyping and genotyping costs, and the accuracy ratio.

This analysis suggests that:

- Even with a lower accuracy, the response obtained with GS will in most cases (provided that $n_0$ is large enough) outperform that of traditional selection (Figure 7A).
- The cost increase (due to genotyping) remains low compared to the traditional cost. This is due to the low genotyping/phenotyping cost ratio. Moreover, the genotyping costs could be compensated by further decreasing the phenotyping costs (e.g. less frequent phenotypic tests, or decreasing the number of replicates) (Figure 7B).
- When the size of the genotyped population increases, the gain increases non-linearly while the cost increases linearly. Thus, the effect on the response of adding more palms to the test population is low above 1000 individuals while the genotyping cost increment remains constant (Figure 7C).
- Consistent with the point above, the increase in response per cost peaks for a fixed population size, which depends on the genotyping and phenotyping costs (Figure 7D).

***Reduction of the genotyping costs***

Since GS requires dense markers at low costs and genomic resources are now available, SNP markers are now favored over other marker types such as SSRs. So far, exploratory studies in oil palm employed SNP genotyping techniques which provide very large numbers of SNPs at relatively high costs (several thousand SNPs, for ~50-300€/palm, which are realistic estimates for the genotyping costs with GBS in Cros et al., 2017a and with the OP300K SNP array in Kwong et al., 2017, respectively). It has been proposed that reducing the number of SNPs used in GS could contribute to reduce the costs. However, this strategy has downsides:

- Reducing the marker number results in the prediction being influenced more by realized relatedness rather than by QTL effects, thereby decreasing the advantage compared to pedigree-based prediction (Jannink et al., 2010).
- Although approaches have been proposed to define the optimal marker sets and these sometimes even improve the prediction accuracy compared to prediction using all available markers (Cros et al., 2017a; Kwong et al., 2017), the defined marker sets are trait-specific, and thus, not necessarily optimal for multi-trait breeding.

In addition to marker number, other technical aspects of genotyping can be optimized. For example, adapting the genotyping technique depending on the number of markers and samples, and minimizing the labor-intensive steps in sample collection and handling.

***Optimization of the GS accuracy***

As illustrated in Figure 7, prediction accuracy is a key factor for GS efficiency. The prediction accuracy reflects how well the model deduced from the training set can predict the genotypic and/or phenotypic value of the tested population. Many factors can affect the accuracy. A brief overview is given below.

❖ Selected trait

Many studies have already highlighted the influence of the trait genetic architecture on GS accuracy. Critical parameters are for example: the QTL number, the heritability, the respective proportion of genetic additivity, dominance and epistasis. Theoretically, the accuracy positively correlates with heritability and this has already been confirmed in oil palm (Kwong et al., 2017) as well as in other species (Covarrubias-Pazaran, 2016; Duangjit et al., 2016; Wolfe et al., 2017). Additive effects are easier to estimate compared to dominant and epistatic effects. Because the part of additivity is generally larger in hybrids, models based solely on additivity can perform well (reviewed in Zhao et al., 2015). In oil palm, this general principle seems to hold true for several yield traits (Cros et al., 2017a; Kwong et al., 2017; Marchal et al., 2016).

❖ Training set

Besides the quality of both genotypic and phenotypic data, the training set design represents a critical factor. The training set combines phenotypic and genotypic data in order to calibrate the model used for prediction. Ideally, the training set should be large and cover all the genetic diversity present in the test population in an unbiased manner (topic reviewed in Zhao et al., 2015). This implies that the relatedness between the training set and the test population must be as high as possible (Cros et al., 2015a; Zhao et al., 2015) while population structure must remain low (exemplified in Duangjit et al., 2016). As a consequence, the training set needs updating along the breeding cycles. Since compliance with these rules can prove difficult when dealing with typical breeding populations, some methods have been proposed to optimize the training set design (Rincent et al., 2017; Wolfe et al., 2017). For breeding companies, a good knowledge of the history and genetics of the breeding population can significantly support the training set design.

❖ Statistical models

A range of statistical methods are available for GS and their efficiency has been already compared in several studies (Covarrubias-Pazaran, 2016; Cros et al., 2015a; Heslot et al., 2012; Jannink et al., 2010; Kwong et al., 2017; Zhao et al., 2015). In many cases, the models display similar performances. On a theoretical point of view, some models might be better suited than others depending on the genetic architecture of the trait considered. Some Bayesian models, for example, can potentially better account for traits which are affected by QTLs with varying effect variances (e.g. a few QTLs with large variance and many with smaller variance), contrary to BLUP which assumes equal effect variance for all QTLs. So far, we have privileged the use of G-BLUP model (which is analogous to rr-BLUP) implemented in ASReml®, since this model has proven its robustness and efficiency for a diversity of trait and species (Covarrubias-Pazaran, 2016; Heslot et al., 2012; Jannink et al., 2010; Zhao et al., 2015), including yield traits in oil palm (Cros et al., 2015a, 2017a).

At the instar of FFB, BN, and ABW in oil palm (Figure 8A-B, Tisné et al., 2015), agronomic traits are often correlated. This implies that independent selection on each correlated trait might not yield the best results. Multivariate GS, together with index selection, has the potential to overcome this issue. Multivariate GS can for example increase the accuracy as shown for BN and ABW in oil palm (Marchal et al., 2016).

❖ Genotypic data

When many markers are available (e.g. several thousand), using all of them can decrease the accuracy and heritability. A similar result is obtained with too few markers (Cros et al., 2017a; Kwong

et al., 2017). The optimal marker number and density is determined by factors such as relatedness, effective population size, and genetic diversity (Jannink et al., 2010; Zhao et al., 2015).

Several studies assessed the impact of marker selection based on criteria such as the marker distribution, LD, and association with the trait. For example, Kwong et al. showed that marker selection based on association and LD can lead to improved accuracy (Kwong et al., 2017). One drawback of this strategy, however, is that the marker set defined is trait- and population-specific.

The quality of the genotypic data can affect the prediction accuracy. For example, missing data is undesirable though imputing can compensate for it, especially when using pedigree data (Cros et al., 2017a). In that respect, genotyping techniques which yield high-quality data at low missing rates should be privileged.

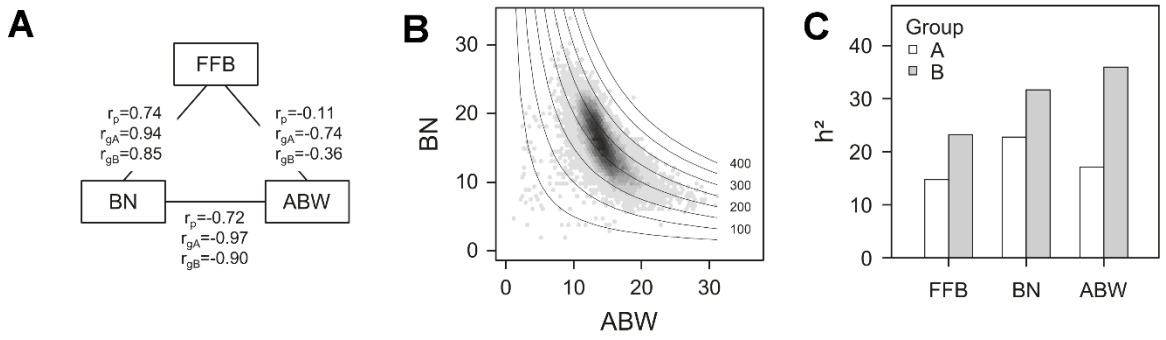



Figure 8 : Heritability and correlations for BN, ABW and FFB depending on the population considered. Figure extracted from Tisné et al., 2015. **A.** Phenotypic correlations ($r_p$) and genotypic correlations in heterotic groups A ($r_{gA}$) and B ($r_{gB}$) between FFB, BN and ABW. **B.** Relationship between average bunch weight (ABW) and bunch number (BN) in a A × B population. The grey scale indicates the density of points with similar BN and ABW values. Isoproduction curves are drawn with corresponding FFB values given on the right of the curves. **C.** Narrow sense heritability ($h^2$) for the three production traits, i.e. FFB, BN, and ABW estimated from A × B individuals.

### *Integrating GS within effective breeding schemes*

How to optimally integrate GS within the selection scheme is a long-standing question. In the context of a breeding program, where resources are limited, implementing GS without any cost increase implies a resource reallocation. Few published studies report investigations on strategies to apply GS for hybrid breeding (Endelman et al., 2014; Longin et al., 2015; Lorenz, 2013; Marulanda et al., 2016; Riedelsheimer and Melchinger, 2013). At least three types of scenarios can be envisaged: within population GS, across population GS, and across generation GS. A comparison of the three scenarios was performed in cassava, which highlights the tradeoff between selection accuracy and distance between training set and test population (Wolfe et al., 2017).

In oil palm, Cros et al. showed the interest of adding GS as a within population pre-selection step to the conventional RRS (Cros et al., 2017a). Though more efficient in terms of genetic gain than classical RRS, the generation time in this scenario does not decrease compared to RRS. In a more recent simulation study, Cros et al. assessed the gain for FFB in breeding strategies where the training set is updated only every second or third generation, and includes individuals from one or two generations (Figure 9, Cros et al., 2017b). As expected, the selection accuracy and gain decreases with the number of generations between GS selection candidates and training set (Figure 10). In this simulation, updating the training set every second cycle by progeny testing and aggregation of data

from two cycles performed best (Figure 10B). Thus, the generation time can be dramatically shortened every second cycle which compensates for the slight decrease in prediction accuracy.

This scheme can certainly be further improved, since many other parameters can be modulated to improve the overall efficiency. The most promising strategies can then be implemented in the field to determine their actual performance.

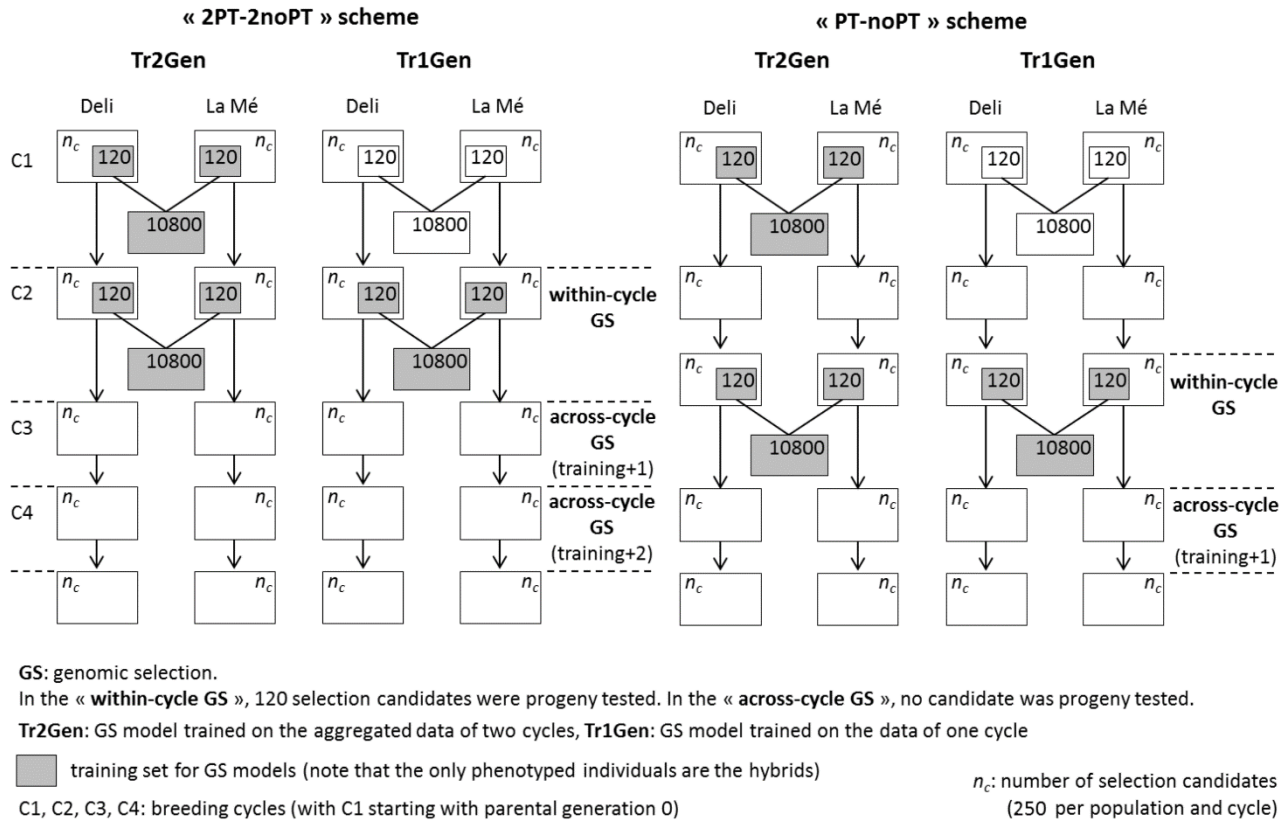



Figure 9: Comparison of breeding strategies involving a training set based on generation (Tr1Gen) or two successive generations (Tr2Gen) and updated every second (PT-noPT) or third generation (2PT-2noPT). 18 individuals were selected within each population at each cycle. Figure from Cros et al., 2017b.

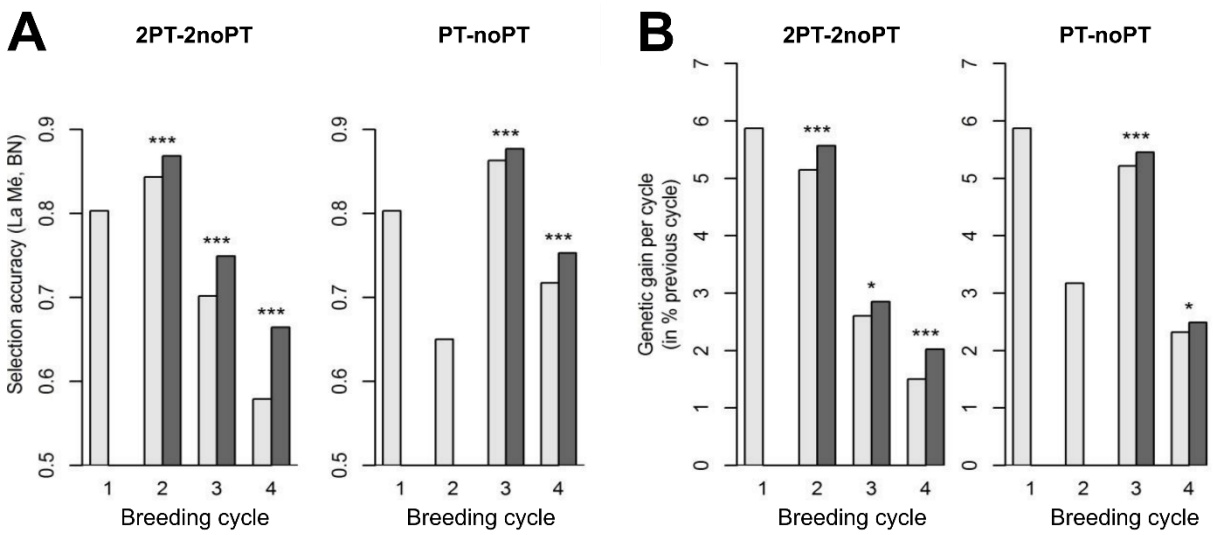



Figure 10: Selection accuracy for BN in group B **(A)** and genetic gain for FFB **(B)** for the breeding schemes described in Figure 9. The breeding population size was fixed to 250 individuals per population and cycle. 18 individuals were selected within each population at each cycle. Data for Tr1Gen and Tr2Gen are presented in light grey and dark grey respectively. Figure from Cros et al., 2017b.

## Conclusion and discussion

### *Concluding remarks*

Similar to what was already shown in other hybrid species (Zhao et al., 2015), GS has the potential to increase the breeding efficiency in oil palm. Simple strategies such as the ones described above could significantly increase the genetic progress in oil palm. Using GS as a pre-selection step can already increase the FFB by 11% over one RRS cycle compared to traditional phenotypic selection (Figure 4, Cros et al., 2017a). Similar to what was observed in black spruce and maritime pine (Bartholomé et al., 2016; Lenz et al., 2017), the selection accuracy is not significantly increased with GS as compared to pedigree-based selection for some agronomic traits (Cros et al., 2017a). For these, a higher gain can only be obtained if GS is associated with a reduction in the generation time and/or an increased selection intensity. So far, oil palm studies have focused on GS for yield traits. However, as suggested by a study in wheat, GS could also be efficient with other traits such as disease resistance (Juliana et al., 2017). This needs to be tested in the future.

Since genotyping generates additional costs, resource reallocation (i.e. by minimizing progeny testing) might be necessary to compensate for those, as proposed in Figure 9. For this, we also concentrate the genotyping effort on the breeding population, thus limiting the number of individuals to genotype, while the value of commercial hybrids (not genotyped) can be accurately predicted based the parents' genetic value. From our own data, we note that the cost of progeny-testing one individual is far above its genotyping cost, and the gap is expected to widen in the future as the genotyping costs are gradually decreasing. The optimal strategy is difficult to determine since many parameters need to be taken into account. Some of them are illustrated in Figure 6. Thus, GS will likely be implemented differently depending on the economic and technical constraints applying to the oil palm breeding companies.

***Challenges and perspectives for GS in oil palm***

Estimating the gain of GS in a simple breeding scenario and for a unique trait is rather simple (e.g. the analysis presented in Figures 6 and 7). Designing and assessing a breeding strategy that integrates GS while allowing efficient selection for all agronomic traits can be more complex, especially when some of these agronomic traits are correlated (example of FFB, BN and ABW in Figure 8A-B). Since the selection strategy needs to take into account the genetics of the breeding population and of the traits under selection, the reality might prove even more complex in oil palm where the A and B breeding populations have fairly distinct history and characteristics, and the agronomic traits have various genetic architectures (Figure 8C and Corley and Tinker, 2015b). Consistent with this, the study by Cros et al. highlights the differential advantage of GS in oil palm depending on the population and the trait (Cros et al., 2017a). Fine tuning will be necessary to develop a breeding strategy which can efficiently select for traits with various genetic architecture and heritability. Since selection on genomic values is relevant for a subset of the agronomic traits only (e.g. quantitative traits), it is then more appropriate to use the term "genomics-assisted selection" to describe the full breeding strategy, which encompasses several other selection steps such as MAS for mono- and oligogenic traits. There are certainly things to learn from other hybrid crop species but breeding in highly valuable perennial crops differs on several aspects from breeding in low-value annual crops. Thus, solutions implemented in other hybrid crops might not be directly applicable to oil palm. In that respect, research in non-hybrid perennial species such as forest trees can be useful (Bartholomé et al., 2016; Isik et al., 2016; Lenz et al., 2017; Resende et al., 2017), even though there is no report on practical GS implementation so far.

On the other hand, there is still some uncertainty about the long-term progress with GS. Reports on GS over many cycles based on empirical data are so far limited. In a multi-parental population of tropical maize, a high genetic gain was achieved using rapid cycling GS over four cycles (Zhang et al., 2017). Predictions can be easily derived from the breeder's formula but this does not take into account genetic drift, random changes in allele frequencies and dominance effects. A loss of genetic variation after each selection cycle is unavoidable. Since the genetic progress correlates with the genetic diversity, the response to selection is expected to decrease over the selection cycles. Thus, managing diversity is a key point for enhancing the long-term progress. By increasing the selection intensity and shortening the breeding cycle, GS could exacerbate this phenomenon, thereby increasing the short-term progress at the expense of the long-term one. This is even more relevant for oil palm where the base population is very narrow.

## Abbreviations

ABW: annual average bunch weight
BLUP: based linear unbiased prediction
BN: annual cumulative bunch number
BV: breeding value
CPO: crude palm oil
FB: fruit-to-bunch ratio
FFB: annual cumulative fresh fruit bunch
G-BLUP: genomic best linear unbiased prediction
GBS: genotyping by sequencing
GCA: general combining ability
GEBV: genomic estimated breeding value
GS: genomic selection
KPO: kernel palm oil
MAS: marker-assisted selection

OER: oil extraction rate
OP: oil-to-pulp ratio
P-BLUP: pedigree-based best linear unbiased prediction
PF: pulp-to-fruit ratio
QTL: quantitative trait locus
rr-BLUP: random regression best linear unbiased prediction
RRS: recurrent reciprocal selection
SCA: specific combining ability
SNP: single nucleotide polymorphism
T-BLUP: traditional best linear unbiased prediction
TS: training set
VS: validation set

## References


Bartholomé, J., Van Heerwaarden, J., Isik, F., Boury, C., Vidal, M., Plomion, C., and Bouffier, L. (2016). Performance of genomic prediction within and across generations in maritime pine. BMC Genomics *17*, 604.

Baudouin, L., Baril, C., Clément-Demange, A., Leroy, T., and Paulin, D. (1997). Recurrent selection of tropical tree crops. Euphytica *96*, 101–114.

Collard, B.C.., and Mackill, D.J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. *363*, 557–572.

Corley, R. h. v., and Tinker, P. b. (2015a). Selection and Breeding. In The Oil Palm, (John Wiley & Sons, Ltd), pp. 138–207.

Corley, R. h. v., and Tinker, P. b. (2015b). The Origin and Development of the Oil Palm Industry. In The Oil Palm, (John Wiley & Sons, Ltd), pp. 1–29.

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLoS ONE *11*.

Cros, D., Denis, M., Sánchez, L., Cochard, B., Flori, A., Durand-Gasselin, T., Nouy, B., Omoré, A., Pomiès, V., Riou, V., et al. (2015a). Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theor. Appl. Genet. *128*, 397–410.

Cros, D., Denis, M., Bouvet, J.-M., and Sánchez, L. (2015b). Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. BMC Genomics *16*, 651.

Cros, D., Bocs, S., Riou, V., Ortega-Abboud, E., Tisné, S., Argout, X., Pomiès, V., Nodichao, L., Lubis, Z., Cochard, B., et al. (2017a). Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. Theor. Appl. Genet. *in press*.

Cros, D., Tchounke, B., and Nkague-Nkamba, L. (2017b). Training genomic selection model across multiple breeding cycles increases genetic gain in oil palm. In Séminaire SelGen 2017, “La Sélection Génomique - Bilan et Perspectives,” (INRA, Paris), p.

Duangjit, J., Causse, M., and Sauvage, C. (2016). Efficiency of genomic selection for tomato fruit quality. Mol. Breed. *36*, 29.

Durand-Gasselin, T., Blangy, L., Picasso, C., Franqueville, H. de, Breton, F., Amblard, P., Cochard, B., Louise, C., and Nouy, B. (2010). Sélection du palmier à huile pour une huile de palme durable et responsabilité sociale. Ol. Corps Gras Lipides *17*, 385–392.

Endelman, J.B., Atlin, G.N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M.E., and Jannink, J.-L. (2014). Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. Crop Sci. *54*, 48–59.

Gallais, A., and Poly, J. (1990). Théorie de la sélection en amélioration des plantes (Masson).

Heslot, N., Yang, H.-P., Sorrells, M.E., and Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. *52*, 146–160.

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C., and Bouffier, L. (2016). Genomic selection in maritime pine. Plant Sci. *242*, 108–119.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics *9*, 166–177.

Juliana, P., Singh, R.P., Singh, P.K., Crossa, J., Huerta-Espino, J., Lan, C., Bhavani, S., Rutkoski, J.E., Poland, J.A., Bergstrom, G.C., et al. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. Theor. Appl. Genet. *130*, 1415–1430.

Kwong, Q.B., Ong, A.L., Teh, C.K., Chew, F.T., Tammi, M., Mayes, S., Kulaveerasingam, H., Yeoh, S.H., Harikrishna, J.A., and Appleton, D.R. (2017). Genomic Selection in Commercial Perennial Crops: Applicability and Improvement in Oil Palm ( Elaeis guineensis Jacq.). Sci. Rep. *7*, 2872.

Lenz, P.R.N., Beaulieu, J., Mansfield, S.D., Clément, S., Desponts, M., and Bousquet, J. (2017). Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC Genomics *18*, 335.

Longin, C.F.H., Mi, X., and Würschum, T. (2015). Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. TAG Theor. Appl. Genet. Theor. Angew. Genet. *128*, 1297–1306.

Lorenz, A.J. (2013). Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. G3 GenesGenomesGenetics *3*, 481–491.

Marchal, A., Legarra, A., Tisné, S., Carasco-Lacombe, C., Manez, A., Suryana, E., Omoré, A., Nouy, B., Durand-Gasselin, T., Sánchez, L., et al. (2016). Multivariate genomic model improves analysis of oil palm (Elaeis guineensis Jacq.) progeny tests. Mol. Breed. *36*, 2.

Marulanda, J.J., Mi, X., Melchinger, A.E., Xu, J.-L., Würschum, T., and Longin, C.F.H. (2016). Optimum breeding strategies using genomic selection for hybrid breeding in wheat, maize, rye, barley, rice and triticale. Theor. Appl. Genet. *129*, 1901–1913.

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.

Rajsic, P., Weersink, A., Navabi, A., and Pauls, K.P. (2016). Economics of genomic selection: the role of prediction accuracy and relative genotyping costs. Euphytica *210*, 259–276.

Resende, R.T., Resende, M.D.V., Silva, F.F., Azevedo, C.F., Takahashi, E.K., Silva-Junior, O.B., and Grattapaglia, D. (2017). Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. Heredity *119*, 245–255.

Riedelsheimer, C., and Melchinger, A.E. (2013). Optimizing the allocation of resources for genomic selection in one breeding cycle. TAG Theor. Appl. Genet. Theor. Angew. Genet. *126*, 2835–2848.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theor. Appl. Genet. 1–17.

Tisné, S., Denis, M., Cros, D., Pomiès, V., Riou, V., Syahputra, I., Omoré, A., Durand-Gasselin, T., Bouvet, J.-M., and Cochard, B. (2015). Mixed model approach for IBD-based QTL mapping in a complex oil palm pedigree. BMC Genomics *16*.

Wolfe, M.D., Carpio, D.P.D., Alabi, O., Egesi, C., Ezenwaka, L.C., Ikeogu, U.N., Kawuki, R.S., Kayondo, I.S., Kulakow, P., Lozano, R., et al. (2017). Prospects for genomic selection in cassava breeding. BioRxiv 108662.

Wong, C.K., and Bernardo, R. (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. TAG Theor. Appl. Genet. Theor. Angew. Genet. *116*, 815–824.

Zhang, X., Pérez-Rodríguez, P., Burgueño, J., Olsen, M., Buckler, E., Atlin, G., Prasanna, B.M., Vargas, M., Vicente, F.S., and Crossa, J. (2017). Rapid Cycling Genomic Selection in a Multi-parental Tropical Maize Population. G3 Genes Genomes Genet. g3.117.043141.

Zhao, Y., Mette, M.F., and Reif, J.C. (2015). Genomic selection in hybrid breeding. Plant Breed. *134*, 1–10.